Xingyang Li

brucelee_sjtu@sjtu.edu.cn / https://acm.sjtu.edu.cn/~xyli

EDUCATION

Massachusetts Institute of Technology — Cambridge, MA
Visiting Student at MIT HAN Lab, EECS Department — 07/2025 – 12/2025
Advisor: Prof. Song Han

Shanghai Jiao Tong University — Shanghai, China
B.Eng. in Computer Science & Technology — 09/2022 – 06/2026 (Expected)
Member of (ACM Honors Class), an elite CS program for top 5% students

- **GPA** (All): **4.03**/4.3, Ranking: **3**/30, 19 A+ courses
- Selected core courses: Computer Vision: 100/100, Machine Learning: 97/100, Comprehensive Design for Computer System: 97/100, Information Theory: 98/100, Advanced Compiler Project: 100/100, Linear Algebra: 100/100

RESEARCH INTERESTS

My research interests lie in efficient machine learning, especially in sparse neural networks and low-bit model quantization. My long-term goal is bringing smaller and more efficient foundation models to everyone as well as making them scalable and accurate.

SELECTED PUBLICATIONS

Radial Attention: $\mathcal{O}(n \log n)$ Sparse Attention with Energy Decay for Long Video Generation
Xingyang Li, Muyang Li*, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, Song Han*
Annual Conference on Neural Information Processing Systems (NeurIPS), 2025. [arxiv] [code] [website] [blog]

Voyager: Real-Time Splatting City-Scale 3D Gaussians on Resource-Constrained Mobile Devices
Zheng Liu, He Zhu*, Xingyang Li, Yirun Wang, Yiming Gan, Wei Li, Yujiao Shi, Jingwen Leng, Minyi Guo, Yu Feng*
Under review at ICLR 2026. [website]

SLTarch: Towards Scalable Point-Based Neural Rendering by Taming Workload Imbalance and Memory Irregularity
Xingyang Li, Jie Jiang*, Yu Feng, Yiming Gan, Jieru Zhao, Zihan Liu, Jingwen Leng, Minyi Guo*
Internation Conference on Computer-Aided Design (ICCAD), 2025; (**Oral Presentation**). [arxiv]

Harnessing Conventional Video Processing Insights for Emerging 3D Video Generation Models: A Comprehensive Attention-aware Way
Tianlang Zhao, Jun Liu*, Xingyang Li*, Li Ding, Jinhao Li, Shuaiheng Li, Jinbo Hu, Guohao Dai*
Design Automation Conference (DAC), 2025; (**Oral Presentation**). [pdf]

RESEARCH EXPERIENCE

Massachusetts Institute of Technology, MIT HAN Lab — Cambridge, MA
Advisor: Prof. Song Han

Efficient and Scalable Video Diffusion Models with Sub-quadratic Complexity — 02/2025 – 07/2025

- We presented the first work of leveraging sparse attention for longer video generation, with sub-quadratic static sparsity mask and minimal LoRA fine-tuning. (*first author, accepted by NeurIPS 2025*)
- I observed the "Spatiotemproal Decay" phenomenon in video diffusion transformer, which states that attention score and density decays as spatial/temporal token distance grow, shedding light on sparsity mask design.
- I designed the $\mathcal{O}(n \log n)$ attention mask for efficient video generation without any tuning, and further extended video length with LoRA fine-tuning.

- Our Radial Attention mask achieved up to 1.9× speedup for default-length video generation, and offered up to 4.4× training cost saving and inference acceleration for up to 4× longer videos.
- Our open-source repository achieves **500+ GitHub Stars**, and has been integrated into the state-of-the-art video diffusion model inference engine.

Shanghai Jiao Tong University, DAI Lab Shanghai, China
Advisor: Prof. Guohao Dai

Mixed Precision Video Diffusion Transformers with Hardware Accelerator 10/2024 – 12/2024

- We presented an algorithm-hardware co-design for efficient video generation models. (*co-first author, accepted by DAC 2025*)
- I designed an attention importance speculation algorithm with a mixed-precision computation strategy for the self-attention operator, which is the bottleneck of video diffusion models.
- We achieved 1.45× speedup and 1.63× energy efficiency compared to existing accelerators.

Shanghai Jiao Tong University, EPCC Lab Shanghai, China
Advisor: Prof. Yu Feng and Prof. Jingwen Leng

Scalable and Low-Latency Mobile Rendering System for 3D Gaussian Splatting 01/2025 – 05/2025

- We presented an effective solution to enable city-scale 3DGS rendering on mobile devices. (*third author, under review at ICLR 2026*)
- I designed an temporal-aware algorithm to reuse level-of-detail tree traversal result with kernel prototypes, alleviating the latency bottleneck on tree traversal stage.
- We achieved over 100× less data transfer and up to 8.9× speedup while retaining rendering quality.

Algorithm-Hardware Co-design for Scalable 3D Gaussian Splatting 08/2024 – 04/2025

- We presented an algorithm-architecture framework for scalable 3DGS, which is bottlenecked by level-of-detail tree traversal and rasterization. (*first author, accepted by ICCAD 2025*)
- I designed an subtree-based tree traversal alorithm with BFS-based re-ordering and the specialized hardware architecture, ensuring streaming DRAM accesses.
- We achieved 3.9× speedup compared to mobile GPU and 1.8× speedup against existing accelerators.

TEACHING EXPERIENCE

Machine Learning 02/2025 – 06/2025
Mathematical Logic 02/2024 – 06/2024
Programming 09/2023 – 02/2024

Role as teaching assistant: giving lectures and recitation classes, writing documents and sample solutions, grading homework, creating exam questions, designing ML-related final project lists, etc.

HONORS

2025 SenseTime Scholarship (**30 winners nationwide** with aspiration in AI research)
2024 Commercial Sponsorship Scholarship (**14 winners** each year in SJTU)
2023 Longfor Merit Scholarship (**Top 10** at Zhiyuan College, SJTU)
2022, 2023, 2024 Zhiyuan Honorary Scholarship (**Top 2%** in SJTU)
2023, 2024 Academic Excellence Scholarship (Ranked 4-th, 2-nd respectively in ACM Class of 2026)

TECHNICAL SKILLS

Language: TOFEL: Total 115 (Reading 30, Listening 30, Speaking 25, Writing 30)
CET 6: 665/710
Programming and Software: Python, C++, CUDA, Bash, Git, Java, Go, and Verilog